DOCUMENT RESUME

ED 173 746 CS 004 712

TITLE	Texas Junior High School Study: Final Report of Process-Outcome Relationships; Classroom Observation Scales. Appendix B.
INSTITUTION	Texas Univ., Austin. Research and Development Center for Teacher Education.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
PUB DATE	Apr 78
CONTRACT	OB-NIE-G-78-0116
NOTE	59p.; R&D Report No. 4061 (COET No. 78-1); For related documents, see CS 004 708 and CS 004 711-713
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Classroom Observation Techniques; *Classroom Research; Elementary Secondary Education; *Measurement Techniques; *Student Behavior; *Teacher Behavior

ABSTRACT

This is the third in a series of four reports describing a study of 1,614 junior high school mathematics and English students and 69 of their teachers that was undertaken to discover the effects of different teaching behaviors on cognitive and affective student outcomes. This booklet contains a discussion of different types of direct observation techniques, including their advantages and disadvantages; 12 scales for use in classroom observation, including definitions of the scales' behaviors and some examples; and a discussion of reliability data for the scales, including between-observer agreement and stability estimates. (TJ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED173746

Texas Junior High School Study:

Final Report of Process-Outcome Relationships

Classroom Observation Scales

Appendix B

R&D Report No. 4061

(COET No. 78-1)

April, 1978

Research and Development Center for Teacher Education
The University of Texas at Austin

This study was supported in part by the National Institute of Education Contract OB-NIE-G-78-0116, The Research and Development Center for Teacher Education. The opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by that office should be inferred.

CS806712

Table of Contents

		Page
	Preface	i
I.	Introduction	1
	Overview of Classroom Observation Procedures	2
	Advantages and Disadvantages of the Four Types of Observation Data	8
	Generating Multi-Point Check Lists	14
	The Set of Classroom Observation Scales	15
	Procedures for Using the Classroom Observation Scales	17
II.	Classroom Observation Scales: The COS System	19
	Level of Attention	19
	Teacher-Initiated Problem Solving	21
	Pupil-to-Pupil Interaction	23
	Teacher Presentation	25
	Negative Affect	26
	Positive Affect	28
	Higher Cognitive Level Student Behavior	30
	Passive Pupil Behavior	35
	Convergent Evaluative Interaction	37
	Rating Scales	39
	Task Orientation	39
	Clarity	41
	Enthusiasm	42
III.	Between Observer Agreement and Stability Estimates for the Scales	43
	Between Observer Agreement	43
	Estimates of Stability	45
	Data Sources	46
	Analyses and Results	46
	Discussion	53
	References	55

Preface

This manual has three parts. The first part is a discussion of different types of direct observation techniques, and of their advantages and disadvantages. Part two presents twelve scales for use in classroom observation. Included are definitions of the scales' behaviors and some examples. The third part consists of the presentation and discussion of reliability data for the scales, including between observer agreement and stability estimates.

The twelve scales presented in this manual are intended for two major uses: to provide measures of classroom process variables for research on teaching and to provide an instructional tool in teacher education. The justification for presenting yet another set of observational measures for these purposes is that the scales provide assessments of classroom behavior on a broader set of variables than the usual direct observation system, they are easy to learn and to use, and they are, for the most part, behaviorally based.

The system is an open-ended one. Other scales may be added or existing scales may be deleted without disturbing the remaining scales. Also, the set of scales can be used to supplement the measures obtained by the kinds of direct observation procedures.

Part I. Introduction

Observation systems may be differentiated from one another on a number of dimensions. Some of these distinguishing characteristics are the unit of observation, such as a time interval or an event; the subject of observation, such as the teacher, pupils, or both; the nature of the behavior which is categorized, for example verbal or nonverbal, or cognitive or affective behavior.

Although few observation systems contain categories of a single type only, usually the predominant number of categories in a system may be classified according to the same set of characteristics. This tends to restrict the number and type of classroom behavior variables that can be measured by any one system. Such a restriction is not a limitation if the variables important to the researcher, evaluator, or teacher are contained within the system. However, the user's interest often lies not only in a few specified variables, but also in obtaining assessments of behavior on a wider range of dimensions than a single observation system allows. For example, a researcher might wish to assess whether some classroom intervention treatment affects variables other than the dependent variables of his study. A supervisor might wish to assess and give feedback about behavior on many aspects of classroom behavior. A teacher educator may be interested in having preservice teachers observe classroom behavior on many dimensions.

To satisfy this need for an assessment of behavior on many dimensions, several alternative procedures can be employed. For example, several

observation systems can be used to obtain the required measurements or a multi-dimensional observation system can be developed to include the appropriate variables. Either of these alternatives is potentially acceptable. However, the observer effort necessary in both instances would be too costly and time-consuming for some applications. For example, few teacher educators will be willing to sacrifice the instructional time needed to make preservice teachers even marginally competent observers on several classroom category systems.

This manual presents another alternative for obtaining data about a variety of dimensions of classroom behavior. This alternative uses a combination of multi-point check list scales and rating scales. Multi-point check lists require an observer to note the frequency of occurrence of a behavior or set of behaviors after a period of observation. A rating scale requires an observer to infer the degree to which some trait or attribute is possessed by the teacher or the pupils. A number of check lists and rating scales will be described, and together they provide a means of assessing classroom behavior along many dimensions. The scales presented in this manual focus upon both the pupil and teacher, verbal and nonverbal interaction, and include both cognitive and affective components. Such diversity in the basis of the observation scales is necessary in order to attempt to capture as much information as possible about classroom processes.

Overview of classroom observation procedures.

Prior to considering the observation procedures presented in this manual, the reader may find it helpful to consider characteristics of observation systems in general. Many procedures are available for observing classroom behavior, and each has particular advantages and disadvantages,

according to the uses to which it will be put.

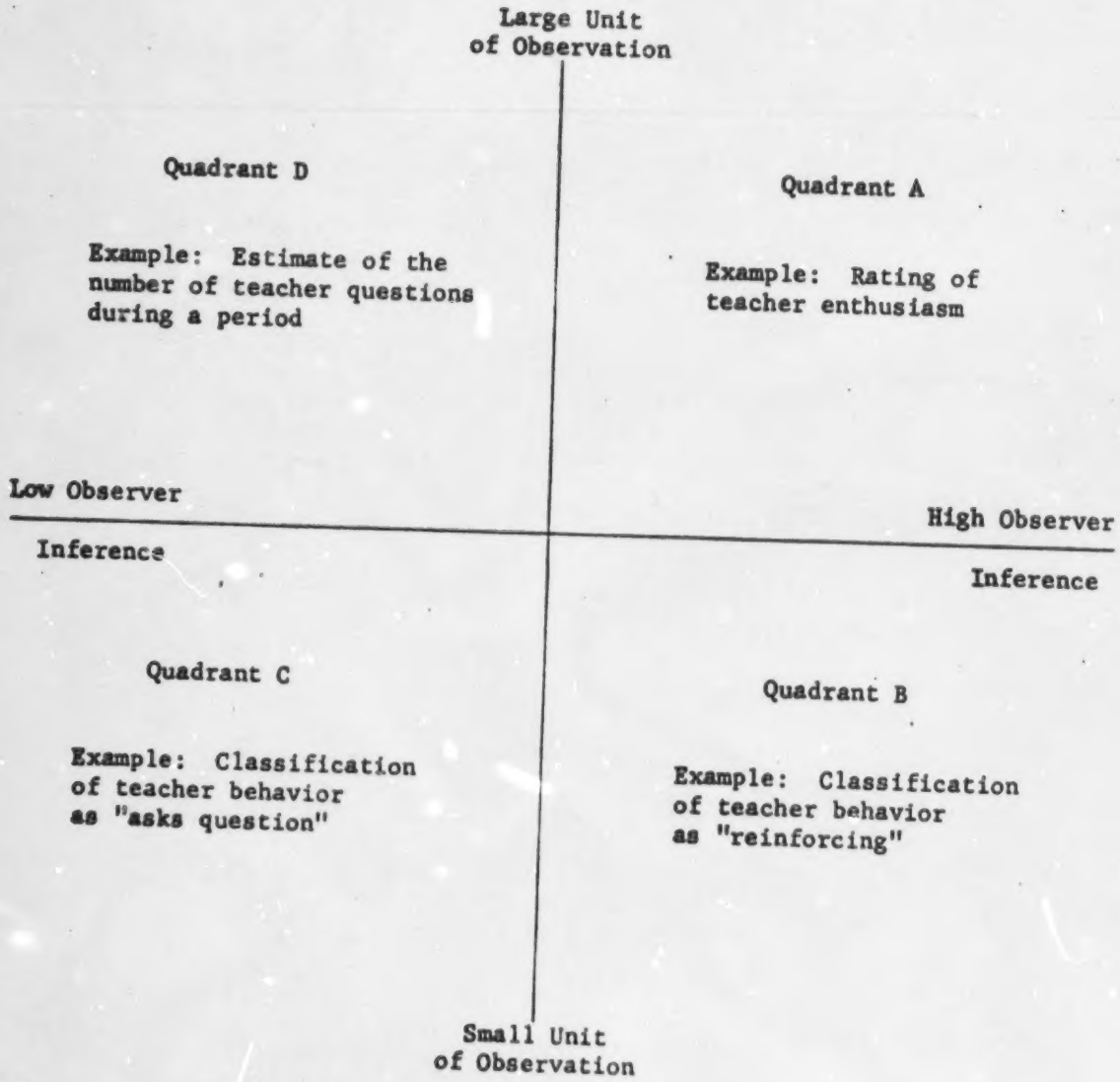
Two ways of classifying observation procedures are especially relevant when one is attempting to construct scales to serve a wide variety of cases. These two characteristics are the observation interval and the degree of inference that is required of the observer.

By observation interval is meant the length of time or the number of events occurring prior to the act of categorization. It is possible to have a very short observation interval, such as three seconds or ten seconds, after which time the observer classifies the behavior that occurred during that time. On the other hand, an observation system may require the observer to classify behavior after a considerable length of time has elapsed or after a large number of events have occurred. Examples of lengthy observation intervals occur whenever the observer is asked to make ratings or classifications after viewing a long segment of instruction, such as a full class period. The second characteristic, degree of observer inference, refers to the extent of interpretation required of the observer. The category "teacher lectures" is a low-inference category because it requires the observer only to note the occurrence or non-occurrence of that event. However, the category "teacher reinforces pupils" requires more interpretation on the part of the observer. In the latter case the observer must somehow judge whether a particular teacher behavior was reinforcing to the student who received it. Although higher inference measurement is often thought of as being unique to rating scales, categories of behavior requiring observer inference are by no means limited to them. The example above of "teacher reinforces pupils" might be found on a direct observation system requiring the observer to categorize each event.

Figure I presents several combinations of observation units and degrees of inference. Beginning in the upper right-hand quadrant (Quadrant A) we find observation categories that require a high degree of inference on the part of the observer and large intervals for units of observation. The most common example of this type of data is a rating scale. An observer may be asked to classify a teacher's warmth or enthusiasm or clarity of presentation or any of a number of such characteristics of either the teacher, the class, or the instruction. Invariably such measurements are made after a period of observation of some length, such as a half hour or 45 minutes. A second type of Quadrant A data is the case study. An observer views classroom behavior for an extended period of time, subsequent to which he writes down his impressions. The impressions usually include judgments of the degree to which the teacher or a pupil possesses a variety of attributes, inferred from their behaviors. Although case studies are not suitable for research purposes, they can serve important functions in the observation component of preservice teacher education.

Quadrant B contains categories that require a high degree of observer inference but whose basic unit of observation is small. Examples may be found in numerous observational category systems. Thus, the following categories, selected from several observation systems, require relatively high degrees of inference: "personal sharing," "positive reinforcement," and "incongruent nonverbal behavior." In each case, the observer must infer some variable or hypothetical construct from an instance of behavior. The inference may be made easier for the observer by providing explicit rules about the conditions governing the inference, but it remains an inference nonetheless.

Figure I. A classification of observation procedures, with examples, according to degree of observer inference and length of observation unit.


Large Unit of Observation
Quadrant D
Example: Estimate of the number of teacher questions during a period
Quadrant A
Example: Rating of teacher enthusiasm
Low Observer Inference
High Observer Inference
Quadrant C
Example: Classification of teacher behavior as "asks question"
Quadrant B
Example: Classification of teacher behavior as "reinforcing"
Small Unit of Observation

The lower left-hand quadrant (C) contains categories having small units of observation and requiring little or no inference on the part of the observer. Examples of such categories are frequent, comprising the major parts of some observation systems. For example, a version of the OScAR system (Medley, et al., 1968) contains many such examples: a category "teacher informs" requires the observer only to judge whether a teacher utterance communicates some content information; the category "pupil questions" requires the observer only to note whether or not the pupil utterance is a question. It is apparent that such categories require a minimum amount of observer interpretation. Since the observer classifies each event as it occurs rather than waiting for several events to occur or for a long time interval to pass, these examples illustrate Quadrant C data. It should be noted that although the observation dimensions in Figure I have been dichotomized, both actually are continuous. There are many, not just two, degrees of inference with respect to observer judgment and interpretation. Also, the length of the observation interval or the number of events prior to classification can be varied. Dichotomization of each continuum has been made for illustrative purposes.

Quadrant D, the upper left-hand area of Figure I, contains categories that require a low amount of inference or interpretation on the part of an observer, but which use large units of observation. Examples of such observation categories may be found among check lists of behavior, requiring the observer only to note the occurrence of an event over a long period of time. For example, the category "Uses audio-visual equipment" is checked by an observer according to whether it has or has not occurred during the period of observation. This is an example of Quadrant D data

because no interpretation other than what is meant by "use" is required of the observer, and because the observation unit is large (the whole period). Notice that if the category were changed to "Uses audio-visual equipment effectively," it would become Quadrant A data, because the judgment of effectiveness requires considerable interpretation, in this case perhaps the prediction of positive consequences on some important pupil outcome variable.

A second kind of low-inference category obtained with large units of observation is a multi-point check list. This type of observation requires an observer to focus upon the occurrence of certain behaviors during a period of time; at the end of the period the observer indicates the frequency of the behaviors or the relative duration of their occurrence. For example, after a half-hour of observation, an observer might indicate on a five-point scale the amount of teacher lecture or informing behavior during that period of time. The points along the scale would be defined according to the frequency of occurrence or relative duration of time occupied by that particular type of behavior. Thus the first point on the scale might indicate zero to 20% lecture, the second point on the scale, 20 to 40% lecture, and so on. This multi-point check list is a low-inference classification procedure because little or no interpretation is required of the observer. What is needed is a reasonably good memory coupled with accuracy in judging the amounts of time or the frequency of events in the period of observation. Again it should be pointed out that the notion of high- versus low-inference categories is an artificial dichotomy, and it is possible to construct multi-point check lists that vary in the degree of interpretation required of the observer.

Advantages and disadvantages of the four types of observation data.

Associated with each of the quadrants are certain advantages and disadvantages with respect to gathering information about classroom behavior and utilizing it in research and in teacher training. These advantages and disadvantages are reviewed below.

Quadrant A data, comprising rating scales and case studies, have the potential for producing data on a large number of dimensions in any one observation. Rating scales may be devised to measure judgments about practically any characteristic of classroom behavior of interest to an investigator. Training time for observers and the level of technical skill needed by an observer to use a set of rating scales is minimal. Furthermore, the use of rating scales does not tax an observer's time or effort. The disadvantages of Quadrant A data are numerous and well-known. The potential for bias because of halo, leniency, and other effects is great. Furthermore, the reproducibility of the behavior used to generate the rating is usually low; consequently it is difficult to use the rating scale directly as a guide for specifying changes in behavior that might produce higher or more appropriate ratings. Also, the advantage listed above of obtaining information about a large number of variables may be illusory. Simply because one defines separate rating scales does not insure that one is measuring separate dimensions of classroom behavior. Without further investigation of the relationships of the rating scales with one another, the user of a system of rating scales has no assurance that he is indeed measuring a large number of classroom behavior dimensions. Nevertheless some of the disadvantages may be overcome through appropriate training of observers and through care in the construction of the system of rating scales.

Quadrant B data, containing high inference categories but utilizing small units of observation, have a number of positive features. Certain of the errors which affect rating scales are of less significance in Quadrant B. For example, since an observer is classifying behavior almost simultaneously with the occurrence of the behavior, there should be no recency or other memory effects. However, the potential for halo, leniency, and other constant errors is still present. The nature of the observation units gives assurance that the scale of measurement will be more precise than is the case with Quadrant A data. That is, Quadrant B data are frequency counts or time durations, having the potential for producing more precise discriminations among teachers or pupils, as compared to five or seven point rating scales. The component behaviors associated with the category of observation are often more identifiable in Quadrant B than in Quadrant A, even though Quadrant B data utilizes high-inference classifications. This happens because the use of Quadrant B data requires an observer to make decisions about behaviors in a very brief period of time. Quadrant A data requires that a summary judgment be made on one occasion. Thus, in order to train observers to obtain Quadrant B data, the component behaviors that cue an observer to classify an event into a Quadrant B category are generally specified in some detail prior to beginning the observations, so the use of Quadrant B data often results in a more explicit statement of the decision rules used by an observer. For example, when training coders to utilize the CASES (Spaulding, 1970) category "pays attention," the characteristics of a child's behavior which denote attentiveness are delimited. The observer must still infer the cognitive state, attentiveness, from the behavior of the child, but at least the

characteristics of the behavior which produce the classification are carefully stated. This process reduces the subjectivity that may enter into the higher inference classification procedure.

The chief disadvantages of obtaining Quadrant B as compared to Quadrant A data usually include longer training time for observers and the likelihood that a limited number of dimensions can be measured by a single Quadrant B observation system. Another difficulty is that the decision rules followed by observers when classifying events into high inference categories may themselves be inaccurate. Because the decision rules are sometimes situationally dependent, being generated within a single study, there is potentially low study-to-study agreement regarding the categories' meanings.

Quadrant C data have several advantages when compared to Quadrants A and B. Because the classifications of behavior in this quadrant are low inference and made after very short durations of time or after single events, there is a higher reproducibility of the events that produce the final scores for an observation. Since little interpretation is required of the observer, the potential for high study-to-study agreement is great. Because the data in this quadrant produce frequency counts of events or durations of time observed in particular categories, the scale of measurement used to discriminate among teachers, classes, or pupils is reasonably precise.

Quadrant C, however, possesses some disadvantages with respect to data collection and interpretation. Usually category systems of this nature require considerable observer training time. While this may not be a disadvantage for research purposes, it is a disadvantage when considering their use in preservice or inservice teacher education. The use of such systems for feedback to teachers often requires considerable training of the teacher

in order to be sure he understands the nature of the observational data. Another drawback to the use of Quadrant C systems in both observational research and teacher training is that the systems typically tap a relatively small number of the potential dimensions of classroom behavior. The fact that a category system may have 20 or 30 categories is no more assurance that the system is measuring a large number of classroom behavior dimensions than is the notion that the use of 20 or 30 rating scales will produce a large number of classroom behavior dimensions. Thus many systems include categories that are primarily indicative of the socio-emotional climate of the classroom, with a few miscellaneous categories to pick up other gross types of behavior. The use of such systems is, of course, appropriate whenever the user is concerned chiefly with those aspects of classroom behavior measured by the observation system. However, many users of observation systems in research and teacher training wish to obtain information about as many dimensions of classroom behavior as is feasible. Although Quadrant C systems very often produce the necessary objectivity and specificity, they do not necessarily provide a diverse set of observable classroom behavior variables.

Quadrant D data use relatively long periods of observation, but require low inference judgments on the part of the observer. If a simple check list is used, the task is to observe classroom behavior for a period of time, after which a judgment is made about whether a given behavior or set of behaviors occurred during the period. If a multi-point check list is used, the observer classifies the behavior or set of behaviors along a scale (perhaps from one to five) indicating the relative amounts of the behaviors in question. To illustrate the distinction between these two types of

Quadrant D data and Quadrant C data, consider the category "Teacher praises pupil." Quadrant C data would be obtained by requiring the observer to code the duration of each occurrence of the behavior. For example, Flanders' system (1965) requires the observer to tally the behavior during three-second intervals. Thus the observer codes the category each time a praise statement occupies the major portion of a three-second interval. In Quadrant D the category "Teacher praises pupil" could be included in two ways. A simple check list might require the observer to note during a half-hour period of observation whether the teacher praised any pupil at all. The categorization would be made whenever a half-hour period occurred during which the teacher emitted at least one praise statement. If no praise statements were forthcoming, the observer would classify that period of observation as having had none. In other words, the scale of measurement for this category of behavior would be reduced to two points. A second type of Quadrant D data could be obtained simply by extending the check list to multi-points, rather than two points. For example, one could define a five-point check list: one end of the scale denotes a low frequency of the behavior and the other end of the scale denotes a relatively higher frequency of this behavior. On a five-point scale, a 1 might indicate the absence of any praise behavior, a 2 might indicate one or two praise statements, a 3 might indicate three or four praise statements, and so on, for a specified duration of observation time.

One disadvantage of Quadrant D data is that the observer's memory will play a considerably greater role in determining the final result in Quadrant D than in Quadrant C. Another problem is the reliance on scales having only a few points. Thus the potential for making precise discriminations among

the subjects of observation is limited. However, if several observations are used for each teacher or class this latter problem will not be as great, since increasing the number of observations has the effect of increasing the precision of the estimate of the true score for the teacher or class.

Quadrant D data has some advantages when compared to data obtained in other quadrants. First, compared to Quadrant A, because the points along the scale of the check list are defined in terms of observable behavior, it is easier to specify the basis for any particular assessment. This specificity should produce reasonably good agreement from study to study among users of such scales. An advantage of Quadrant D data, which it shares with Quadrant A, is that the coding process is simpler than is the case with Quadrant B or Quadrant C data. For example, to use the scales presented in this manual, the observer need only record his observations at 15-minute intervals and only a brief time is actually needed to code. Naturally the observer cannot be asleep during the other parts of that 15-minute period; he must attend to the behaviors of the teacher and pupils. However, the observer does not need to engage in an elaborate coding process. The ease with which Quadrant D data may be gathered allows the inclusion of a large number of variables into an observation system. In addition, a variable may be added or removed without disturbing the categorization of behavior into other scales or without requiring redefinitions of remaining scales, as might be the case with Quadrants B and C. Because the coding procedure is simple, training time is minimized. However, coders must still learn the definitions for the behaviors and the various scale points.

The use of Quadrant D observation scales for assessment and feedback to teachers may also have some advantages. First, the feedback recipient

does not need to learn very much about the observation procedure in order to understand it. Second, since Quadrant D procedures, and particularly multi-point check lists, can be used to gather data ranging across a large number of classroom behavior variables, they offer the potential for considerable flexibility and adaptibility to a variety of classroom settings and teaching concerns. Finally, Quadrant D procedures may prove valuable in preservice teacher education, for some of the reasons given above. In addition, most teacher education programs include some observation experiences in their curriculum activities. Students may be taught the use of Quadrant D scales as a part of this experience. Thus their attention could be focused upon a wide range of classroom behaviors without resorting to elaborate training, coding and data tabulation procedures, while still retaining much of the objectivity and reproducibility of direct observation systems.

Generating multi-point checklists.

The development of multi-point check lists can proceed in a number of ways. Scales can be generated a priori, according to the purposes of the study or other use, or existing classroom observation systems may be utilized by selecting from them those behavior categories that determine the dimensions of interest. One may then define Quadrant D scales that measure analogous dimensions, by defining points along the scales according to the frequency or duration of the selected categories of behavior. The selection procedure used to generate the observation scales can be empirical or logical. That is, if one selects several existing observation systems, categories from them may be combined either through logical or empirical procedures to form variables. One need not necessarily combine categories, but it is generally more parsimonious to do so. Building scales non-empirically requires developing

a rationale for combining categories based upon psychological, curricular, or other bases. An empirical process bases combinations upon the interrelationships among category scores. Of course, the latter procedure requires data.

An empirical procedure (factor analysis) was used as the basis for some of the observation scales presented in this manual. Specifically, observations were collected on five occasions in 28 fifth- and eighth-grade classrooms using five observation systems: OScAR5 (Medley, et al., 1968), FAIR (Fuller, 1970), SES (Spaulding, 1970), Dyadic (Good and Brophy, 1970), and CCS (Emmer and Albrecht, 1970). Intercorrelations among categories in each of these systems were factor analyzed, and resulting factors were intercorrelated across systems to produce common factors. A complete description of the factors obtained in the study is presented in Emmer and Peck (in press). To form some of the check lists presented in this manual, a number of the factors from that study were selected, and scales were defined utilizing the behavior categories associated with each factor. Thus some of the scales presented in this manual are composites of several behavior categories that were correlated with the factor represented by the scale. That is, the scales represent clusters of categories describing behavior of teachers and pupils.

The set of classroom observation scales.

The scales described in this manual are listed in Figure 2.

I. Pupil attention
II. Teacher-initiated problem solving
III. Pupil-to-pupil interaction
IV. Teacher presentation
V. Negative affect
VI. Positive affect
VII. Higher level cognitive pupil behavior
VIII. Passive pupil behavior
IX. Convergent-evaluative interaction
X. Task orientation
XI. Clarity
XII. Enthusiasm

Figure II. Classroom Observation Scales

Scales II, III, V, VI, VII, VIII, and IX were based upon the factors obtained from a single system or a combination of systems, as described above. Scales I and IV are single category scales that originally were poles of separate bipolar factors. Because the opposite ends of those factors were not well defined, only one end of each factor was used to define the scale. The three final scales, X, XI, and XII were not empirically based. Instead, they were added because they measure variables relating (although not necessarily causally) to positive pupil outcomes in several different research studies. In addition, the latter scales tap aspects of classroom processes not directly assessed by the other scales.

In the terms of the scheme presented earlier for describing types of observation systems, scales X, XI, and XII (task orientation, clarity, and enthusiasm) are the purest examples of Quadrant A data. Scale I is also a Quadrant A type, although it is a little less subjective and thus is somewhat closer to Quadrant D. Although the observer must still make an inference from observed behavior, nevertheless the criteria for such an inference are reasonably explicit. The remaining scales fall within Quadrant D; that is,

they require little or no observer inference and are coded after a fairly long period of observation.

Procedures for using the Classroom Observation Scales.

To use the Classroom Observation Scales, the observer should sit so he can observe both the teacher and the students. If possible, he should be in a position to observe the facial features of as many students as possible. Then, at 15-minute intervals, the observed behaviors are coded onto a form, such as the one shown in Figure III.

I.	ATTN	(H)5	4	3	2	1(L)
II.	TIPS	(H)5	4	3	2	1(L)
III.	P-PINT	(H)5	4	3	2	1(L)
IV.	% T PRES	(H)5	4	3	2	1(L)
V.	Neg Af	(H)5	4	3	2	1(L)
VI.	Pos Af	(H)5	4	3	2	1(L)
VII.	COG.	(H)5	4	3	2	1(L)
VIII.	PASS	(H)5	4	3	2	1(L)
IX.	CVG-EV	(H)5	4	3	2	1(L)
X.	TSK OR	(H)5	4	3	2	1(L)
XI.	CL	(H)5	4	3	2	1(L)
XII.	ENT	(H)5	4	3	2	1(L)

Figure III. Classroom Observation Scales Sample Form

The left hand column contains abbreviations for the scales. The "H" and "L" in parenthesis remind the coder of the high and low ends of the scales. Six or eight of these forms can be comfortably spaced on one side of an 8 ½ x 11" page.

In the early stages of coding, observers find it helpful to carry with them the definitions of the scales to refer to.

The recommended coding interval is 15 minutes. Shorter intervals are not likely to produce enough behaviors on some of the scales to effectively

discriminate among classes or teachers. Longer intervals may reduce accuracy by placing too much reliance on the observer's memory. Fifteen minutes is an arbitrary figure, however, and users of these scales may wish to alter the time interval to suit their special conditions.

Part II. Classroom Observation Scales: The COS System

In this section each of the twelve scales is described in detail. For each scale, its defining behavior or set of behaviors is first described. Then, the definitions of the various points along the scale are presented, followed by more detailed definitions and examples of the behaviors, when necessary. Finally, ground rules or special circumstances affecting use of the scale are included in the presentation.

I. Level of Attention

Attention as defined for this scale refers to pupil orientation toward the teacher, the task at hand, or whatever classroom activities are appropriate. If a pupil is attending to inappropriate activities, or is engaged in self-directed behavior when he is supposed to be engaged in a class activity, his behavior is not considered attentive. Therefore, you should look for behavior that is focused upon or engaged in whatever activity is appropriate, be it individual seat work, group discussion, or listening to the presentation of information. Cues that are useful in recognizing attentive behavior include eye contact, body orientation, response to questions or other eliciting cues, and participation in class activity. Cues that are useful in recognizing inattentiveness include inappropriate social interaction, repetitive body movement, visual wandering, and engagement in behavior other than the sanctioned class activity. At times, it will be difficult to determine whether a student is being attentive, such as when the teacher presents information and the student sits facing the teacher, with no observable behaviors indicating inattention. In such an instance, the pupil is considered attentive until he provides a behavioral indicator to the contrary.

To code behavior on this scale, scan the class on several occasions during each 15-minute observation period. Note how many pupils appear attentive or inattentive at a given time. After some practice you will find it easier to keep in mind your estimate of the number of inattentive students. Average your estimates of attention and record it using the scale below.

1. Fewer than half of the students are attentive most of the time.
2. One-half to three-fourths of the students appear attentive most of the time; the remainder are attentive only some of the time.
3. Most of the students are attentive, but several (four to six) are attentive only some of the time.
4. Nearly all students are attentive, but a few (one, two or three) are attentive only some of the time.
5. All the students are attentive most of the time.

Note -- the phrase "most of the time" means at least 75% of the time the observer checks the pupils for attentiveness.

Some examples from which inattention and attention may be inferred:

Inattention: Moving around the room at inappropriate times.

Reading a book during class discussion.

Two boys whispering.

Child sits with elbows on desk, her fingers holding her eyelids open.

Doodling with a pencil.

Child lays her head on her desk.

Child asks a question unrelated to the activity of the class.

Staring fixedly at an object not related to class activity.

Attention: Raising hand to volunteer a response.

Maintaining eye contact, following teacher's movements.

Turning to watch another child who is contributing to the class activity.

Working on assigned activity.

If a free activity period, pupil is engaged in some task.

II. Teacher-Initiated Problem Solving

This scale indicates the degree to which the teacher exhibits a particular style of instruction. Specifically, teacher-initiated problem solving refers to a pattern of behavior in which the teacher frequently addresses questions and problems to the entire class. The student response to the question is often followed up by the teacher's directing subsequent elaborating questions to other students. Frequently the teacher probes or delves into student answers. Whether there is much pupil initiated interaction is not relevant to this scale, and it may in fact be low.

To use this scale you will have to judge the extent to which the teacher's behavior during the 15-minute period of observation can be characterized by the above description.

1. Little or no teacher-initiated problem solving. Other activities or teacher behaviors predominate.
2. Occasionally there is teacher-initiated problem solving, but most interaction would be characterized by something else (e.g. lecture, recitation, drill).
3. Some of the behavior is teacher-initiated problem solving. Some questions and problems are addressed to the class, but not frequently; the teacher occasionally probes pupils' responses,

and asks elaborating questions.

4. Much of the teacher's behavior is teacher-initiated problem solving, but other types of teacher behavior (e.g. presenting information, giving directions) occur frequently -- as much as one-half of all teacher behavior.

5. Most of the teacher's behavior is teacher-initiated problem solving.

Because this scale is a composite of several behaviors, examples and definitions of each are presented below.

Teacher questions or problems directed to the class:

A substantive utterance in question form phrased by the teacher and directed at the class is considered a problem. The only exceptions are those questions that occur during drill activities, requiring only rote response from the students. It is important to note that this does not include a question directed at a specific student; rather, the whole group is addressed with the question prior to selecting a student to respond.

Teacher (to class): When you got off the Mayflower, what was it you decided to do before you came off? John?

John: I think we had to make some rules.

The question is directed to the class before designating John as the respondent. If the teacher named John to answer prior to asking the question, then the question would not be directed to the class, and would not qualify as a behavior indicative of the teacher-initiated problem solving scale.

Example: Teacher sets up class in two debating groups, and then interjects questions and problems for them to consider.

Example: Teacher: Are all people alike? I think I'm exactly the same as you, am I?

Example: Teacher: What is a subset of the set of real numbers?

Teacher probes, delves into student answers:

The teacher asks a subsequent question of the student to elicit clarification, elaboration, explanation, or further ideas.

Example: Student: I think we had to make some rules.

Teacher to Student: Why did you say it was necessary to make more rules? (Asks for explanation.)

Teacher to Student: Were there any other things you did after getting off the Mayflower? (Asks for other ideas.)

Teacher to Student: How necessary was it to make rules? (Asks for elaboration.)

Teacher to Student: What kinds of rules do you mean? (Asks for clarification or elaboration.)

Teacher asks elaborating questions:

Elaborating questions are those which are addressed to pupils other than the one who gave the preceding answer. These questions build upon or make some use of the previous response.

III. Pupil-to-Pupil Interaction

Substantive pupil utterances in which one pupil interacts with another pupil, a group of pupils, or responds indirectly to the teacher are classified on this scale. These behaviors may be observed in a variety of activities. For example, the pupils may be interacting in groups, or students may react verbally to the utterances of other students during a discussion. Pupil behavior on this scale also includes responding indirectly to the teacher. This means that a pupil answers a teacher's question without being called upon to do so. When this occurs, it usually follows another pupils' response to the teacher and is at least in part a reaction to the first pupil's comment or response. Generally, the teacher's behavior during

periods having high pupil-to-pupil interaction will be non-evaluative; that is, there will be little or no criticism, praise, or approval. To code, note approximately how much of the 15-minute observation is spent in the behaviors that characterize this scale.

1. No pupil-to-pupil interaction. All pupil behavior involves routine responding, following directions, individual seat work, etc.
2. One or two instances of pupil-to-pupil interaction.
3. Some instances of these behaviors but not much. Major portion of student behavior is routine responding, seatwork, listening to the teacher, etc.
4. Many of these behaviors, but not as much as one-half of the period is taken up by them.
5. Student to student interaction occurs frequently during one-half or more of the observation period.

Examples:

Pupil-to-Pupil interaction:

Pupils work in groups on a problem assigned to them by the teacher, and the interaction in groups relates to the assigned work. If this activity occurs for one-half or more of the period, a 5 would be coded. Note that the extent to which the activity has been structured by the teacher is not relevant to this scale.

Indirect pupil response to the teacher:

Teacher: How many beans are in this pot? Give an estimate.

Pupil 1: About a thousand.

Pupil 2: That's too much. A hundred. (indirect)

Pupil 3: In between a hundred and a hundred and ten. (indirect)

Teacher: Why is atomic power important? John?

John: Because it's used to help man.

Pupil 2: Well, we can use it to generate electricity.

Pupil 2's behavior is an indirect response since the teacher did not call on him to respond. John's behavior is a direct response to the teacher's question. The observer must decide whether the teacher designates the pupil, verbally or non-verbally, to respond; an indirect response is coded only when the pupil is not designated to respond. The only exception occurs when pupils are interacting and the teacher indicates who is to respond next; that is, the teacher is performing a gatekeeping function. For example, if Pupil 2 reacts to Pupil 1's response, and the teacher merely indicated that Pupil 2 could respond, then Pupil 2's reaction is considered indirect. Note -- Chorus responding is not considered an instance of pupil-to-pupil interaction. Pupil-to-pupil interaction is coded only when it is substantive; that is, refers to content-relevant topics.

IV. Teacher Presentation

This scale measures only one type of behavior. The observer's task is to estimate the relative amount of class time occupied by teacher presentation of substantive information. By teacher presentation is meant substantive (content oriented) verbal or non-verbal behavior that provides information, and does not imply or require pupil response, nor evaluate pupil behavior. Thus, teacher questions, procedural directions, praise, and criticism are not instances of teacher presentation. Lecture, reading to the class, answering pupil questions, and any other activity in which the teacher gives information will be an instance of teacher presentation.

To use this scale, observe the teacher's behavior and note the amount of teacher presentation as compared to the total of all teacher behavior, pupil behavior, and periods of seat work or other activities in which there is no verbal interaction.

1. Teacher presentation occurs 0-20% of the period.
2. Teacher presentation occurs 20-40% of the period.
3. Teacher presentation occurs 40-60% of the period.
4. Teacher presentation occurs 60-80% of the period.
5. Teacher presentation occurs 80-100% of the period.

Note that in order for a 5 to be scored, the observation period must be taken up almost entirely by teacher information giving. Even a small amount of discussion or other activity will be likely to cause the rating to be a 4 or less. On the other hand, a score of 1 will occur only when there is a very small amount of teacher information giving (less than 20% of the time).

Teacher presentation should not be confused with teacher talk, since the former may be only a small part of the latter. Although a question and answer session might contain 70% teacher talk, much of it may be teacher questions and evaluation rather than presentation. Be sure to distinguish teacher presentation from other teacher behaviors.

V. Negative Affect

This scale includes behaviors that show negative or hostile feelings on the part of either or both the teacher and pupils. The behaviors that form this scale include teacher criticism or correction, teacher desist statements, teacher sarcasm or ridicule, pupil hostility, and pupil resistance. The behaviors may be both verbal and nonverbal.

To use this scale, note the frequency and intensity of instances of the negative affect behaviors that occur during the 15-minute observation period.

1. No negative affect.
2. One instance of mildly negative behavior.
3. Two or three mildly negative behaviors.
4. Several mildly negative behaviors, or one or two moderately negative behaviors.
5. Several moderately negative instances or one or more instances of severe negative affect.

Examples:

Teacher criticism: verbal or nonverbal disapproval of pupil behavior.

Teacher (to student): Don't be so sloppy in your work.

Teacher (to class): If you don't behave more like adults, you aren't going to go on the field trip.

Teacher (to pupil): Either participate like the rest of the class or leave the room.

Desist statements: Telling pupils to stop doing something.

Teacher (in irritated voice): OK, let's all close our lips. Stop talking.

Teacher (to pupil): Stop bothering her.

Other negative behavior: Includes nonverbal behaviors, such as shaking of head to indicate disapproval, pointing, hitting, or other negative physical contact. Some other verbal behavior includes asking questions which call attention to inappropriate pupil behavior (e.g. "Why are you doing that?" or "Can't you listen just once?"), or calling a pupil's name to get his attention and distract him from inappropriate behavior.

Pupil hostility: Includes negative pupil behavior directed at the teacher or at another student. Examples of such behavior are name-calling, sarcasm, criticism, belligerence and talking back.

Pupil resisting: Passive aggressive behavior, including deliberately ignoring instructions, foot-dragging, or refusal to work or participate.

The intensity of the behaviors also will influence the final rating. A single mildly negative behavior only indicates a rating of 2, whereas a single severely negative instance will cause a 5 to be rated. In judging intensity, the observer should attend to the emotionality displayed by the participants and other pupils' reactions to the event.

Mild intensity: Very little emotion displayed by the participants or the teacher; other pupils unaffected.

Moderate intensity: Some emotion evident (tenseness, sharp tone of voice, or raised voice) and other pupils notice the event or seem a little affected by it.

Severe intensity: Considerable emotion displayed (may involve physical contact, reddening of face, or a loud voice), the event captures the attention of or disrupts the non-participants' activities.

Note -- As an aid to the interpretation of scores on this scale, the observer can indicate whether the rating is due to the teacher's interaction with a single student, or whether several students or the whole class were involved.

VI. Positive Affect

This dimension comprises those teacher behaviors that show support of and positive regard for pupils and their behavior. Such acts include smiling, kidding and joking, encouragement and hints, praise, moving close to pupils (in a positive way), touching affectionately, and making supportive

statements. Observe the teacher's behavior during each 15-minute segment and note the frequency of these behaviors. As with negative teacher affect, the intensity of the behavior will influence the score assigned to the teacher's behavior. Thus, the use of "Good" as a response to a correct answer, if it has become perfunctory or insincere, does not qualify for a high rating even if it is repeated a dozen times.

1. Absence of positive affect (but not necessarily negative). No hints or encouragement.
2. Some positive affect, but its expression is limited to rather perfunctory acceptance of pupil answers.
3. Occasional smiles, hints or praise by teacher, but not particularly intense.
4. One or two instances of obviously sincere support or regard, accompanied by other teacher approach behavior such as smiling, encouraging or praising.
5. Three or more instances of obviously sincere support or regard. Teacher smiles at students, approaches students, provides praise.

Most observers have little trouble recognizing positive affect when it occurs. The examples below are intended to convey the different levels of intensity which distinguish "obviously sincere" support or regard from more moderate expression of support and praise.

Examples of strong positive affect behaviors:

Teacher: It's Columbus' birthday today and it's also Junior's birthday, so let's sing "Happy Birthday" to him. (T brings Junior to front of class and puts her arm around him while the class sings "Happy Birthday.")

Teacher (to class): You're really smart today. You've got your thinking caps on.

Teacher (to Jimmy): You're really thinking today, Jimmy. Good for you!

Teacher (to pupil): You really gave a good report today. (Teacher pats student on shoulder.) That shows a lot of good effort.

Examples of behaviors indicative of more moderate affect:

Teacher: OK, you're on the right track.

Teacher: That was a good answer.

Teacher: Did you hear the last answer? That was good.

Also included are such behaviors as nodding the head to indicate approval of a response, and such expressions as "OK," "Right," "Fine," and "Go on." Such responses are indicative of at least some support or regard if they are uttered with some inflection or emphasis. However, if they have become an habitual teacher response to pupil behavior, and are devoid of any intensity, then they should be discounted as indicative of positive affect.

You should also note that a score of 1 means that there was no evidence of positive affect; it does not necessarily imply the presence of negative feelings or behavior.

VII. Higher Cognitive Level Student Behavior

This scale focuses upon pupil verbal behaviors that give evidence of higher or more complex cognitive processes. Higher cognitive processes are involved when a student makes a generalization or inference, explains an an answer by citing data or rules, solves a problem by combining or using other principles or rules, and defines concepts by citing classes of objects

or events, rather than single examples. Student behavior not indicating higher cognitive processes includes rote or simple associative reproduction of responses, defining concepts in terms of a single attribute or example, stating impressions and opinions without justification, and simple descriptive reporting of events or observations with no attempt to explain or draw inferences from them.

The behaviors on this scale, more than any other, will occur as a function of the age of the pupils being observed. Higher cognitive level pupil behavior, as defined on this scale, will not be observed very frequently in most first- or second-grade classrooms. In high-school classes, such behavior may occur frequently during class discussions. The scale presented here is intended for use in intermediate grades; in other settings users may wish to alter the definitions for the various scale points to encompass the range of such behaviors observed in that setting.

It is very important to note that this scale is intended to measure only the degree of cognition present in the verbal utterances of pupils. Consequently, the observer should not try to judge whether a student is "really thinking" or whether the unstated basis for a "yes" or "no" response to a teacher's question is actually a generalization or an inference. The observer simply is trying to assess the degree to which higher cognition is present in the overt behavior of the class. A dozen super intellects might be clicking and whirring covertly during a class discussion, but unless their behavior is verbalized, this scale is not intended to assess it. To use the scale, note the frequency of verbal behaviors you would classify as exhibiting higher cognitive processes. Do not attempt to classify nonverbal behavior, such as seatwork, boardwork, etc. Then record your impression on the scale below.

1. Higher processes do not occur.
2. Higher processes occur rarely (one instance only, or perhaps several questionable occurrences).
3. Higher processes occur occasionally (two or three reasonably certain instances).
4. Higher processes occur moderately (four or five certain or reasonably certain instances).
5. Higher processes occur frequently (six or more certain instances).

Definition and examples:

Generalization: A rule, principle, or conclusion based upon specific instances or data.

Example: After examining results of a study, a student concludes that "environment influences personality."

Example: In an arithmetic class, generating a new rule to solve a problem not previously encountered, such as discovering that division may be checked by taking the product of the divisor and quotient to obtain the dividend.

Explanation: To invoke a principle or rule to justify an answer.

Example: Teacher: How do you know that the two expressions will still be equivalent?

Student: Because of the commutative property of multiplication.

Example: Teacher: Why didn't more people cross the ocean?

Pupil 1: Because of monsters.

Teacher: Well, how did it happen that they got that idea, when we know that there probably weren't any?

Pupil 2: One person tells another, and then others tell others,

and pretty soon so many people are talking about it that people believe it.

Note -- The response by Pupil 1 would not be considered explanation because no rule or principle was cited, only a single event (or object). If Pupil 1 had explained why monsters kept people from crossing the ocean, then he would be credited with a higher process response. Pupil 2 gets credit for an explanation because his response states a principle, in this case a principle of behavior. Notice that sophistication, correctness, or phraseology are not relevant to the classification.

Example: Teacher (presents picture of oil well): Now look at the picture; where do you think this oil well is located?

Pupil 1: Israel.

Pupil 2: California.

Pupil 3: New Mexico.

Pupil 4: Oklahoma.

Pupil 5: El Paso.

Teacher: Why El Paso?

Pupil 5: Well, look at the background. There are Yucca and it looks like a desert; there are tall, brown mountains, like around El Paso.

The response by Pupil 5 is explanation because the inference was made based upon observed data. The other responses are not considered as higher level cognition because no reasons are stated. This illustrates the rule concerning the use of this scale to classify a response as higher level cognition: the reasons, explanation, generalization, etc. must be verbalized. Otherwise, raters' estimates will not agree with each other because of different subjective criteria that will be employed.

Defining Concepts: Stating the attributes of a concept; however, presenting a single example of a concept with no further explanation is considered description and not higher level cognition, as defined by this scale. Concepts may be concrete or abstract, simple or complex.

Example: Teacher: What makes something alive?

Pupil 1: Horses are alive.

Pupil 2: It grows, lives, reproduces, and dies, like a plant.

Note -- Pupil 1 has not provided any of the attributes, only an example, so the response does not meet the criterion for defining concepts. Pupil 2 lists several attributes along with the example, so the response is at a higher level.

Example: Teacher: What is a role?

Pupil: Well, I'm a member of my family and a member of the Boy Scouts, and I act one way with my family and another way at Scouts.

Note -- This is considered higher cognitive level because one attribute of the concept is presented, even though its form may be rudimentary.

Distinguishing higher cognitive level behavior from other types of behavior will be easier if you can keep in mind the definitions and examples of lower level cognitive behavior presented below.

Rote, associative responding: Verbatim reproduction of verbal sequences. A stimulus-response chain without apparent mediators.

Examples: Pupil responses during drill activities.

Reading from a book or list.

Responses to flash cards.

Description: Stating facts or describing the appearance, characteristics or physical attributes of an object or event, with no attempt to explain, define, or draw conclusions.

Examples:

Teacher: What is the climate of Oregon like?

Pupil: It's often damp, very wet climate, and the mountains on the west produce rainfall.

Teacher: Name at least three purchases the U.S. made during the period of expansion.

Pupil: The Louisiana Purchase, The Gadsen Purchase, and Alaska.

Teacher: What's the highest mountain in the world?

Pupil: Mount Everest.

Teacher: What's another word for terrain?

Pupil: Land.

VIII. Passive Pupil Behavior

This scale measures the extent to which pupils engage in passive, as opposed to active, modes of behavior. Withdrawal by the pupil from engagement with his surroundings is classified as passive. In addition, visual wandering and passive observation in which the student avoids maintaining contact for any length of time is considered passive behavior.

To use this scale, scan the class several times during your observation, noting the frequency with which pupils exhibit passive behavior at the various times.

1. No more than one of the pupils is passively disengaged at any one time.

2. A few pupils appear passive occasionally.
3. Several pupils appear passive occasionally or a few pupils are passive much of the time.
4. As many as five or six students are passive much of the time.
5. Passive behavior exhibited by a third or more of the class, much of the time.

Definitions and examples of passive behavior:

Withdrawing behavior: Behavior whose effect is to disengage the student from participation in his environment.

Examples: Pupil does not respond when called upon.

Daydreaming. Staring fixedly at some object or out the window.

Repetitive movements with no apparent function, such as finger tapping, rocking, or swaying.

Pupil lays his head down on his desk, or he slouches down in his seat with his eyes closed.

Visual wandering, passive observation: Looking about with no interest or attempt to engage in interaction.

Examples: Pupil's eyes move from window, to teacher, to hall, to other pupils, etc. He does not follow the main pattern of interaction in the class.

Pupil ignores lesson or seatwork and watches two other pupils conversing.

The key to using the scale is the idea of passive disengagement from ongoing classroom activities. Pupils may be actively disengaged as well, for example, by inappropriate social interaction or self-directed activity.

These latter behaviors do not reflect passive student behavior, however, even though they are disengaged.

To some extent, the scale is related to the attention scale. Pupils who are coded as very attentive will not be observed as passive, since attention implies engagement in the ongoing activity. Nevertheless, the Passive Pupil Behavior scale provides additional information in many cases. For example, a class observed as relatively inattentive and having a 1 or 2 on passivity is quite different in behavior from a class with the same level of attentiveness and with a 4 or 5 rating on passive behavior. Except for classes with very high attentiveness scores, passivity scores can vary across the full range.

IX. Convergent-Evaluative Interaction

This scale measures a particular style of interaction. It is characterized by a focus upon obtaining the correct answer to the teacher's question, with little or no attempt to continue the contact once the answer has been obtained. The following behaviors are typical of this style: teacher questions, usually having an identifiable, correct answer (as opposed to divergent questions), are accompanied by pupil response to the questions. If the pupil response is incorrect, the teacher corrects it, and sometimes probes the response when she is not satisfied with it. If the pupil response is correct, the teacher will usually accept the response, acknowledging its correctness, but will not probe to find the basis for the answer.

To use this scale, estimate the relative amount of all interaction that you consider to be convergent-evaluative. If there is no interaction (e.g. all lecture or all seatwork), then do not use the scale.

1. Very little or no convergent-evaluative interaction; all or almost all behavior follows some other pattern.
2. Some of the interaction is convergent-evaluative but most interaction is something else.
3. Up to one-half of the interaction is convergent-evaluative.
4. Much of the interaction is convergent-evaluative.
5. Almost all interaction is convergent-evaluative (75% or more).

Examples of convergent-evaluative interaction sequences:

Teacher: Name the parts of the cell.

Pupil: Membrane, nucleus, protoplasm, cell wall.

Teacher: Correct.

Teacher: What would a state like Oklahoma have to import?

Pupil: Things they don't have very much of.

Teacher: OK. Can anyone give an example of what Oklahoma would import?

Pupil 2: Lumber.

Teacher: That's right. What else?

Pupil 3: Vegetables from the Gulf.

Teacher: The Gulf?

Pupil 3: I mean the Valley.

Teacher: OK. Anything else?

Pupil 4: Hardware from the New England states.

Teacher: Would you write the answer to the problem on the board? (Student writes on board.)

Teacher: That's correct. What's the answer to number six?

Pupil 1: Eight.

Teacher: No, your mistake was understandable, but the answer is ten. What's the answer to number seven, Jon?

Pupil 2: Twelve.

Teacher: Good.

Rating Scales

The final three observation scales are somewhat different than those presented previously. The earlier scales require one to look for specific behaviors or sets of behaviors, and then to estimate their frequency of occurrence. The last three scales require more inferential judgments on the part of the observer. Even so, we have found that observers are able to use these scales quite accurately. These scales are included because they are attempts to assess variables that have been found in at least some empirical studies to be related to positive pupil outcomes. Moreover, these three scales assess aspects of teaching not directly measured in the preceding nine scales. Thus, they provide potentially useful and non-redundant information about classroom processes.

X. Task Orientation

This scale is a measure of the degree to which the teacher works toward content-related, substantive goals. Low task orientation occurs when class sessions place little emphasis upon student attainment of content objectives, as indicated by wandering discussions, wasting time with procedural activities, busy-work assignments, and unproductive activity. High task orientation occurs when there is a definite emphasis upon student attainment of content objectives, when class activities appear to be carefully planned, and when discussions are substantive and focused.

1. Very low task orientation. Much wasted time, pointless activities or discussion, and little substantive progress attempted by the teacher.
2. Low task orientation, but neither very low nor moderate.
3. Moderate task orientation. Attainment of content objectives is usually, but not always, the main purpose of class activities; class activities and procedures are occasionally a waste of time and effort.
4. High task orientation, neither moderate nor very high.
5. Very high task orientation. Emphasis clearly is upon student attainment of content objectives; class activities appear to be carefully planned; discussions are fruitful; a minimum of time is lost to procedural activity.

In using this scale, you will find it helpful to adopt the perspective that it is primarily the teacher's task orientation you are assessing, and only secondarily, the class'. Time wasted by the students when they have been given an assignment by the teacher does not count for a low rating, although failing to help the students get on with a task does count for a low rating. The latter rating depends upon the observer's judgment that the need for such help was or should have been apparent to the teacher. Assignments given just to keep pupils busy (admittedly, a subjective judgment) get lower ratings than assignments that relate clearly to the content objectives that are the focus of the lesson. Remember however, that you are not observing a Marine Corps drill team or a new-car assembly line, so a certain amount of wasted time and motion is to be expected in even the most task-oriented teacher's classroom or learning center.

XI. Clarity

Clarity refers to the degree to which the teacher's presentation of material and his substantive interactions with students are understood by them. Low clarity means that the teacher is "over their heads" and is confusing to the pupils. High clarity means that the pupils appear to understand the teacher. The observer must use many cues to make his rating, noting particularly the pupils' reactions to the teacher's presentation and instructions as well as the teacher's actual behavior. When observing the teacher for clarity, check for behaviors such as stating the objectives of the activity or assignment, giving directions that do not require constant repetition, giving several examples or using different media to illustrate a point, checking with students at various times to verify their understanding, and summarizing main points from discussions or lessons. Other than the absence of the above characteristics of clarity, indicators of low clarity include looks of bewilderment on the faces of the pupils, statements that they do not understand or that they are confused, frequent incorrect pupil responses to the teacher's questions, teacher statements that use qualifiers (e.g. maybe, sometimes, not always) without seeking or supplying the salient qualifying attributes, and vague or evading teacher responses to pupil questions.

1. Very low clarity. Pupils seem very confused by the presentation. The teacher cannot answer the pupils' questions, or answers them in an unclear manner by using concepts and terms the pupils are apparently unfamiliar with or by being overly complex and ambiguous.
2. Low clarity. Between very low and moderate.
3. Moderate clarity. The teacher seems to be understood by most pupils,

but not all of the time. Sometimes the teacher is confusing and vague.

4. High clarity. Between moderate and very high.
5. Very high clarity. The teacher's explanations are easy to understand and pupil questions are adequately answered. The teacher seems aware of the pupils' levels, sensing problems they are having or may have.

XII. Enthusiasm

This scale is used to judge the extent to which the teacher displays interest, vitality, and involvement in his subject and his instruction.

1. Very low enthusiasm. The teacher's behavior is lethargic, dull, routine; a minimum of vocal inflection, gesturing, movement, or change in facial features. The teacher appears to lack interest in what he is doing.
2. Low enthusiasm. Between very low and moderate.
3. Moderate enthusiasm. Occasionally the teacher seems interested and involved; some display of activity, such as gesturing. Sometimes the teacher is dull, routine, and lacking in vigor.
4. High enthusiasm. Between moderate and very high.
5. Very high enthusiasm. The teacher is stimulating, energetic, and very alert. He seems interested and involved in what he is teaching; moves around, gestures, inflects voice.

Part III. Between-Observer Agreement and Stability Estimates for the Scales

The scales presented in this manual have been used on several occasions by observers in research studies. This section presents some data bearing upon the problems of the accuracy of observers and the stability of the characteristics measured by the scales. The observers who collected the data discussed below received approximately six hours of training, including practice observations using both video tapes and live classrooms. A few of the observers had teaching experience, but most had not; both graduate and undergraduate students participated as observers.

Between-observer agreement

The question of the extent to which observers agree upon the classification of behaviors into certain categories is important when considering the use of the scales for any purpose. Between-observer agreement indicates whether the coders are assigning similar scores to the same behaviors. This is the only way it can be known whether the same criteria are being used by the coders in applying the scales and whether the scales have been clearly defined.

Two indices of agreement are present in the tables: the intraclass correlation and the average difference score. The intraclass correlation is obtained from a comparison involving the amount of variation among observers and the amount of variation among classes or teachers. If there is as much variation on a scale among observers, averaged across observations, as among classes or teachers; then there is no basis for assuming that observers agree beyond what would be expected by chance. High coefficients indicate that most of the variability among scores is caused by differences among teachers or classes, rather than by disagreement among coders.

The average difference score is computed by subtracting an observer's score on a scale from the average of all other scores on the scale for the same observation. This procedure is repeated for each observer, and the absolute value of each difference score thus obtained is calculated and then averaged across observers. For example, suppose three observers code using the scale "Pupil Attention." The first observer codes a 3 and the second and third observers each code 4. The difference score for observer one is +1, for observer two, -.5, and for observer three, -.5. The average difference score for the observation is therefore .67, since $|+1| + |-.5| + |-.5| = 2.0$ and $2.0/3 = .67$. The average difference scores reported in the tables are based upon averaging these difference scores across observations, thus they are an estimate of how much difference to expect between any two observers' ratings on any observation.

The average difference score is presented in addition to the intraclass correlation for a heuristic reason. The intraclass correlation does not provide an estimate of agreement in terms of the original units of measurement, whereas the average difference score does. For example, a coefficient of observer agreement of .60 can be interpreted in terms of explained and unexplained variance, but not directly in terms of how far apart two observers' ratings are on the original scale. There is not a one-to-one correspondence between the intraclass correlations and the average difference scores presented in the tables because the intraclass correlation depends in part upon the variation in scores among classes or teachers, and this is not reflected directly by the average difference score.

The observations used here for data sources were longer than 15 minutes, and therefore required several codings per observation. When computing

the average difference score with multiple codings per observation, the average difference between observers was computed for each coding, and then averaged across codings. Thus the average difference score still reflects the difference between observers for any given 15-minute coding period. When computing the intraclass correlation between observers on such data, the scores for an observer were averaged across codings prior to computation of the coefficient for each scale.

Estimates of stability

The stability of a scale is an important characteristic when considering the scale's use for research or evaluation. High stability means that, for a particular teacher or class, observations made at different times will yield similar scores. Low stability means that a score obtained in an observation made on one occasion does not predict very well the score obtained on another occasion.

The estimates of stability presented in this section are based upon the intraclass correlation which also was used for between-observer agreement. However, instead of comparing variation among observers, the stability coefficients are based upon a comparison involving the variation in scores for several observations per class. High coefficients indicate that much of the variation in scores can be explained by systematic differences among teachers or classes. The higher the intraclass correlation for a scale, the greater is its stability, and consequently its power to discriminate among teachers or classes.

Data sources

Data set one: The sources for this data set were video tapes obtained in 15 fifth-grade classrooms, with five observations occurring in each class. Technical failures and other problems reduced the number of available tapes from 75 to 68. The tapes were each one-half hour long, with the content of the lessons ranging across arithmetic, social studies, and science lessons. Two observers coded each video tape independently. Since the tapes were one-half hour in length, each observation is made up of the average of the scale scores for the two 15-minute coding periods.

Data set two: The sources for this data set[1] were observations made by ten observers in 31 first- and second-grade classrooms. In most cases, four observations were made in each class, usually with two observers present. Observers in this study generally coded the COS from two to four times during the period of observation.

Analyses and results

Between-observer agreement for data set one was based upon 68 observations in 15 classrooms, with two observers per observation. Table 1 lists the between-observer agreement estimates based upon the intraclass correlation. Two sets of correlations are presented for each scale. The first column estimates the reliability (agreement) of a single observer's score; the second column is the estimate of the reliability of the average of two observers' scores.

[1]Data set two was obtained in a research study directed by Dr. Jere Brophy. The author is grateful to Dr. Brophy for making these data available.

Between-observer agreement for data set two was obtained in the following manner. One observation was sampled randomly per teacher (N=31), and the two observers' estimates for that observation were used in the computations. The intraclass correlations are listed in Table 2, with the estimate of the reliability of a single observer in the first column and the reliability of the average of two observers in column two. In both tables 1 and 2 the probabilities are based upon a null hypothesis of no agreement between observers.

Table 3 contains the average difference score for each scale in data sets one and two.

Statistics (intraclass correlations) pertaining to the stability of each of the twelve COS scales are shown in Tables 4 and 5. Table 4 lists stability estimates based upon data set one. The first column gives the stability of a single observation while the second column estimates the stability of the average of five observations of a class. The third column contains the probabilities associated with a test of the null hypothesis of no stability among observations.

Table 5, which presents stability data from data set two, is similar to Table 4, except that column two gives the stability of the average of four, rather than five, observations per class.

TABLE 1

Between-Observer Agreement for Data Set One

Scale	ρ_1	ρ_2	p<
1. Pupil Attention	.45	.62	.001
2. Teacher-Initiated Problem Solving	.46	.63	.001
3. Pupil-to-Pupil Interaction	.53	.69	.001
4. % Teacher Presentation	.71	.83	.001
5. Negative Affect	.50	.67	.001
6. Positive Affect	.47	.64	.001
7. Higher Level Cognitive Pupil Behavior	.60	.75	.001
8. Passive Pupil Behavior	.21	.35	.10
9. Convergent Evaluative Interaction	.70	.82,	.001
10. Task Orientation	.07	.13	N.S.
11. Clarity	.39	.56	.001
12. Enthusiasm	.35	.52	.01

TABLE 2

Between-Observer Agreement for Data Set Two

Scale	ρ_1	ρ_2	P<
1. Pupil Attention	.81	.89	.001
2. Teacher-Initiated Problem Solving	.57	.73	.001
3. Pupil-to-Pupil Interaction	.77	.87	.001
4. % Teacher Presentation	.45	.62	.01
5. Negative Affect	.78	.88	.001
6. Positive Affect	.68	.81	.001
7. Higher Level Cognitive Pupil Behavior	.06	.12	N.S.
8. Passive Pupil Behavior	.53	.69	.001
9. Convergent Evaluative Interaction	.33	.48	.05
10. Task Orientation	.70	.82	.001
11. Clarity	.42	.60	.01
12. Enthusiasm	.51	.68	.01

TABLE 3

Average Difference Scores for Pairs of Observers for Data Set One ($\bar{D}_1$) and Data Set Two ($\bar{D}_2$)

Scale	$\bar{D}_1$	$\bar{D}_2$
1. Pupil Attention	.31	.59
2. Teacher-Initiated Problem Solving	.63	.90
3. Pupil-to-Pupil Interaction	.90	.74
4. % Teacher Presentation	.49	.59
5. Negative Affect	.53	.75
6. Positive Affect	.72	.73
7. Higher Level Cognitive Pupil Behavior	.68	.80
8. Passive Pupil Behavior	.38	.85
9. Convergent Evaluative Interaction	.70	.67
10. Task Orientation	.54	.66
11. Clarity	.38	.68
12. Enthusiasm	.65	.59

TABLE 4

Stability Estimates From Data Set One

Scale	ρ_1	ρ_5	R_K
1. Pupil Attention	.36	.75	.001
2. Teacher-Initiated Problem Solving	.48	.81	.001
3. Pupil-to-Pupil Interaction	.28	.64	.01
4. % Teacher Presentation	.44	.78	.001
5. Negative Affect	.54	.84	.001
6. Positive Affect	.53	.83	.001
7. Higher Level Cognitive Pupil Behavior	.28	.63	.01
8. Passive Pupil Behavior	.25	.60	.01
9. Convergent Evaluative Interaction	.38	.73	.001
10. Task Orientation	.08	.27	N.S.
11. Clarity	.27	.62	.01
12. Enthusiasm	.65	.89	.001

TABLE 5

Stability Estimates From Data Set Two

Scales	ρ_1	ρ_4	p<
1. Pupil Attention	.64	.87	.001
2. Teacher-Initiated Problem Solving	.33	.66	.001
3. Pupil-to-Pupil Interaction	.46	.77	.001
4. % Teacher Presentation	.29	.61	.001
5. Negative Affect	.21	.51	.01
6. Positive Affect	.42	.74	.001
7. Higher Level Cognitive Pupil Behavior	.32	.64	.001
8. Passive Pupil Behavior	.25	.57	.01
9. Convergent Evaluative Interaction	.10	.30	.10
10. Task Orientation	.45	.76	.001
11. Clarity	.58	.84	.001
12. Enthusiasm	.52	.81	.001

Discussion

From the between-observer agreement estimates in Tables 1 and 2 it can be observed that most of the 12 COS measures were used with at least moderate accuracy by single observers in both studies. Thus, it may be inferred that the descriptions of the scales are sufficiently precise to be used appropriately and consistently by different observers. Two scales, Higher Level Cognitive Pupil Behavior and Task Orientation, were unreliable in one of the two studies, while Passive Pupil Behavior (8) had only marginally significant agreement in one study. However, each of these scales evidenced adequate agreement in the other set of data. Users of these scales should therefore be especially watchful during observer training for inaccuracies in these scales.

Table 3 contains some implications for observer training. Average difference scores greater than one do not occur in either set of data. Thus, while not necessarily precluding significant agreement (since this depends in part upon sample variation on the scale) average observer differences greater than one during training would be a signal that agreement is likely to be poor. Differences less than one do not however guarantee good agreement. Still, the rule of thumb: "Average differences greater than one are likely indicators of inaccuracy," provides a good basis for early checks of observer agreement during training.

The estimates of stability in Tables 4 and 5 have several implications. First, on only a few of the scales was the stability of a single observation sufficiently high to serve as the basis for predicting future performance with any degree of accuracy. Thus, the use of these scales in research, evaluation, or feedback is very limited if only one observation is to be

made. However, when four or five observations are made, and the average score on a scale is used as the basic datum, then the stability of the scales increases greatly in most instances. For two or three observations, the reliability of a scale will of course lie in between the two estimates provided in the tables. The reader can determine the values by applying the Spearman-Brown prophecy formula to the "test length," etc. With four or more observations the scales generally appear to attain quite adequate reliability for most research purposes.

On the basis of the observer agreement and the stability results, the 12 COS scales appear suitable for use in preservice teacher education. Some potential applications are acquainting the student with a wide array of instructional variables, increasing his awareness of classroom processes, and providing him with behaviorally-based concepts for thinking and conversing about classroom behavior.

The use of these scales for feedback or self-assessment, e.g., during student teaching, should be limited to situations in which more than one observation-assessment is made. Otherwise the user runs the clear risk of making judgements based on data that are not sufficiently reliable. However, the fact that with several observations most of the scales attain reasonably good reliability indicates that they can be used to help a teacher acquire a fairly accurate assessment of his teaching on the measured characteristics.

References

Brophy, J.E. and Good, T.L. Teacher-Child Dyadic Interaction: A Manual for Coding Classroom Behavior. The Research and Development Center for Teacher Education, The University of Texas at Austin, 1970.

Emmer, E.T. and Albrecht, I.D. The Cognitive Components System Manual. The Research and Development Center for Teacher Education, The University of Texas at Austin, 1970.

Emmer, E.T. and Peck, R.F. Dimensions of Classroom Behavior. Journal of Educational Psychology (in press).

Flanders, N.A. Teacher Influence, Pupil Attitudes, and Achievement. Office of Education, Cooperative Research Monograph No. 12, 1965.

Fuller, F.F. FAIR System Manual. The Research and Development Center for Teacher Education, The University of Texas at Austin, 1970.

Medley, D.M., Schluck, C.G. and Ames, N.P. Assessing the Learning Environment in the Classroom: A Manual for Users of OScAR 5V. Princeton, New Jersey: Educational Testing Service, 1968.

Spaulding, R.L. Classroom Behavior Analysis and Treatment. San Jose State College. November, 1970.